

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re U.S. Patent Application of)
TANAKA et al.)
Application Number: To Be Assigned)
Filed: Concurrently Herewith)
For: FABRIC AND METHOD FOR SHARING AN I/O)
DEVICE AMONG VIRTUAL MACHINES FORMED IN)
A COMPUTER SYSTEM)
ATTORNEY DOCKET NO. GOTO.0007)

Honorable Assistant Commissioner
for Patents
Washington, D.C. 20231

**REQUEST FOR PRIORITY
UNDER 35 U.S.C. § 119
AND THE INTERNATIONAL CONVENTION**

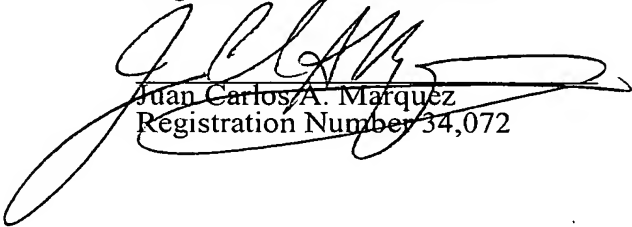
Sir:

In the matter of the above-captioned application for a United States patent, notice is hereby given that the Applicant claims the priority date of February 18, 2003, the filing date of the corresponding Japanese patent application 2003-040232.

The certified copy of corresponding Japanese patent application 2003-040232 is being submitted herewith. Acknowledgment of receipt of the certified copies is respectfully requested in due course.

Respectfully submitted,

Stanley P. Fisher
Registration Number 24,344


Juan Carlos A. Marquez
Registration Number 34,072

REED SMITH LLP
3110 Fairview Park Drive
Suite 1400
Falls Church, Virginia 22042
(703) 641-4200
December 5, 2003

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 3 年 2 月 1 8 日
Date of Application:

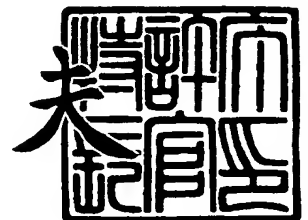
出 願 番 号 特 願 2 0 0 3 - 0 4 0 2 3 2
Application Number:
[ST. 10/C] : [J P 2 0 0 3 - 0 4 0 2 3 2]

出 願 人 株 式 会 社 日 立 製 作 所
Applicant(s):

2 0 0 3 年 9 月 2 6 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康



出証番号 出証特 2 0 0 3 - 3 0 7 9 3 9 2

【書類名】 特許願

【整理番号】 GM0212045

【提出日】 平成15年 2月18日

【あて先】 特許庁長官殿

【国際特許分類】 G06F 17/30

【発明者】

 【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所 中央研究所内

 【氏名】 田中 剛

【発明者】

 【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所 中央研究所内

 【氏名】 上原 敬太郎

【発明者】

 【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所 中央研究所内

 【氏名】 對馬 雄次

【発明者】

 【住所又は居所】 東京都国分寺市東恋ヶ窪一丁目 2 8 0 番地 株式会社日立製作所 中央研究所内

 【氏名】 濱中 直樹

【発明者】

 【住所又は居所】 神奈川県秦野市堀山下 1 番地 株式会社日立製作所 エンタープライズサーバ事業部内

 【氏名】 吉田 大輔

【発明者】

 【住所又は居所】 神奈川県海老名市下今泉 8 1 0 番地 株式会社日立製作所 インターネットプラットフォーム事業部内

 【氏名】 若井 義憲

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100075513

【弁理士】

【氏名又は名称】 後藤 政喜

【選任した代理人】

【識別番号】 100084537

【弁理士】

【氏名又は名称】 松田 嘉夫

【選任した代理人】

【識別番号】 100114236

【弁理士】

【氏名又は名称】 藤井 正弘

【手数料の表示】

【予納台帳番号】 019839

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 0110326

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 計算機システム、I/Oデバイス及びI/Oデバイスの仮想共有方法

【特許請求の範囲】

【請求項1】

計算機の制御プログラム上で構築された複数の仮想計算機と、

前記計算機のP C Iバスに接続されるとともに、前記複数の仮想計算機で共有されるI/Oデバイスと、を備えた計算機システムにおいて、

前記I/Oデバイスに配設されてP C Iバスに接続された単一のポートと、

前記複数の仮想計算機のうちのひとつと、前記ポートとの論理的な接続状態を設定するP C I接続割り当て手段と、

前記仮想計算機からの制御信号に基づいて、前記P C I接続割り当て手段に設定された接続状態を更新するI/Oデバイス切り換え手段とを備え、

前記仮想計算機は、P C I接続割り当て手段の設定に基づいてI/Oデバイスの変更を行うことを特徴とする計算機システム。

【請求項2】

前記I/Oデバイス切り換え手段は、前記P C I接続割り当て手段の設定を更新するとともに、仮想計算機にI/Oデバイスの変更を通知する割り込みをかける割り込み手段と、を有し、

前記割り込みを受けた仮想計算機は、前記P C I接続割り当て手段の設定に基づいてI/Oデバイスの変更を行うことを特徴とする請求項1に記載の計算機システム。

【請求項3】

前記仮想計算機は、他の仮想計算機に障害が発生したことを検知する障害検知手段を有し、障害を検知したときには前記I/Oデバイス切り換え手段に予め設定した制御信号を送出することを特徴とする請求項1または請求項2に記載の計算機システム。

【請求項4】

前記仮想計算機は、第1の仮想計算機と第2の仮想計算機を含み、

前記第2の仮想計算機は、第1の仮想計算機に障害を検知したときには前記I/Oデバイス切り換え手段に予め設定した制御信号を送出し、I/Oデバイスのポートを第2の仮想計算機へ接続するとともに、前記制御プログラムは、第2の仮想計算機を稼働させる一方、第1の仮想計算機を待機させることを特徴とする請求項2または請求項3に記載の計算機システム。

【請求項5】

計算機を物理的に分割した複数の物理分割計算機と、

前記計算機のPCIバスに接続されるとともに、前記複数の物理分割計算機で共有されるI/Oデバイスと、を備えた計算機システムにおいて、

前記I/Oデバイスに配設されてPCIバスに接続された単一のポートと、

前記複数の物理分割計算機のうちのひとつと、前記ポートとの論理的な接続状態を設定するPCI接続割り当て手段と、

前記物理分割計算機からの制御信号に基づいて、前記PCI接続割り当て手段に設定された接続状態を更新するI/Oデバイス切り換え手段とを備え、

前記物理分割計算機は、PCI接続割り当て手段の設定に基づいてI/Oデバイスの変更を行うことを特徴とする計算機システム。

【請求項6】

前記I/Oデバイス切り換え手段は、前記PCI接続割り当て手段の設定を更新するとともに、物理分割計算機にI/Oデバイスの変更を通知する割り込みをかける割り込み手段と、を有し、

前記割り込みを受けた物理分割計算機は、前記PCI接続割り当て手段の設定に基づいてI/Oデバイスの変更を行うことを特徴とする請求項5に記載の計算機システム。

【請求項7】

前記計算機は、複数の物理分割計算機に障害が発生したことを検知する障害検知手段を有し、障害を検知したときには前記I/Oデバイス切り換え手段に予め設定した制御信号を送出することを特徴とする請求項5または請求項6に記載の計算機システム。

【請求項8】

前記物理分割計算機は、第1の物理分割計算機と第2の物理分割計算機を含み

、
前記障害検知手段は、第1の物理分割計算機に障害を検知したときには前記 I/O デバイス切り換え手段に予め設定した制御信号を送出し、I/O デバイスのポートを第2の物理分割計算機へ接続するとともに、前記計算機は、第2の仮想計算機を稼働させる一方、第1の仮想計算機を待機させることを特徴とする請求項6または請求項7に記載の計算機システム。

【請求項9】

計算機の P C I バスに接続される I/O デバイスにおいて、
前記 I/O デバイスは P C I バスに接続される単一のポートと、
計算機からの制御信号に応じて、前記ポートの論理的な接続状態を変更する割り込み信号を発生する信号発生手段とを備え、
前記計算機は、この割り込み信号があったときに前記ポートの論理的な接続状態を変更することを特徴とする I/O デバイス。

【請求項10】

前記計算機は、第1の計算機と第2の計算機を含み、
前記信号発生手段は、第1の計算機からの制御信号に基づいて、第2の計算機へ割り込み信号を送り、前記ポートの論理的な接続を前記第1の計算機に切り換えることを特徴とする請求項10に記載の I/O デバイス。

【請求項11】

前記信号発生手段は、割り込み信号を発生するとともに、前記ポートの論理的な接続状態を設定する割り当て手段を更新することを特徴とする請求項9または請求項10に記載の I/O デバイス。

【請求項12】

計算機の P C I バスに接続された I/O デバイスを、計算機の制御プログラム上で構築された複数の仮想計算機で共有する I/O デバイスの仮想共有方法において、

前記 I/O デバイスは、単一のポートを介して P C I バスに接続されて、前記複数の仮想計算機のうちのひとつと、前記ポートとの論理的な接続状態を設定する

手順と、

前記仮想計算機からの制御信号に基づいて、前記ポートと仮想計算機との論理的な接続状態を切り換える手順と、を含むことを特徴とする I/O デバイスの仮想共有方法。

【請求項 13】

前記接続状態を切り換える手順は、前記ポートと仮想計算機との論理的な接続状態の設定を変更するとともに、前記仮想計算機に I/O デバイスの変更を通知する割り込みをかける手順と、

前記割り込みを受けた仮想計算機が、前記論理的な接続状態の設定に基づいて I/O デバイスの変更を行うことを特徴とする請求項 12 に記載の I/O デバイスの仮想共有方法。

【請求項 14】

前記接続状態を切り換える手順は、前記複数の仮想計算機のいずれかに障害が発生したことを検知したときに、前記ポートと仮想計算機の論理的な接続状態を設定する割り当て表を更新するとともに、障害が生じた仮想計算機を待機させるとともに、他の仮想計算機を稼働させることを特徴とする請求項 12 または請求項 13 に記載の I/O デバイスの仮想共有方法。

【請求項 15】

計算機の PCI バスに接続された I/O デバイスを、計算機を物理的に分割した複数の物理分割計算機で共有する I/O デバイスの仮想共有方法において、

前記 I/O デバイスは、単一のポートを介して PCI バスに接続されて、前記複数の物理分割計算機の一つと、前記ポートとの論理的な接続状態を設定する手順と、

前記物理分割計算機からの制御信号に基づいて、前記ポートとの論理的な接続状態を切り換える手順と、を含むことを特徴とする I/O デバイスの仮想共有方法。

【請求項 16】

前記接続状態を切り換える手順は、前記複数の物理分割計算機の内いずれかに障害が発生したことを検知したときに、前記ポートと物理分割計算機の論理的な接

続状態を設定する割り当て表を更新するとともに、障害が生じた物理分割計算機を待機させるとともに、他の物理分割計算機を稼動させることを特徴とする請求項 15 に記載の I/O デバイスの仮想共有方法。

【請求項 17】

前記接続状態を切り換える手順は、前記複数の物理分割計算機のいずれかに障害が発生したことを検知したときに、前記ポートと物理分割計算機の論理的な接続状態を設定する割り当て表を更新するとともに、障害が生じた物理分割計算機を待機させるとともに、他の物理分割計算機を稼動させることを特徴とする請求項 15 または請求項 16 に記載の I/O デバイスの仮想共有方法。

【請求項 18】

計算機の P C I バスに接続した I/O デバイスを複数の計算機で共有する I/O デバイスの仮想共有方法において、

前記 I/O デバイスは単一のポートを介して P C I バスに接続され、複数の計算機のいずれかひとつからの制御信号に応じて、前記ポートの論理的な接続状態を変更する割り込み信号を発生する手順と、

この割り込み信号に基づいて、前記ポートと計算機の論理的な接続状態を変更する手順と、を含むことを特徴とする I/O デバイスの仮想共有方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、P C I デバイスのホットプラグが可能な O S が動作する計算機システムに係り、特に、制御プログラムによって P C I デバイスを論理的に挿入、除去操作してホットプラグ操作をする計算機システムに関する。

【0002】

【従来の技術】

計算機システムで 24 時間 365 日サービスといった無停止長時間運用を実現する上で、いかに耐障害性を高めるかが問題となっている。J I M ・ G R A Y 著「トランザクション処理 概念と技法」によれば、近年はソフトウェア障害の割合が増大している。このことから、ソフトウェア障害に対処することが特に必要

となっているといえる。

【0003】

ソフトウェア障害の原因として、OSやアプリケーションソフトが処理のために占有したメモリ領域を開放しないまま放置してしまうために起きるメモリリークやアプリケーションソフトのバグなどがある。このような障害に対処するために、計算機システムの管理ソフトウェアで定期的にサービスやOSの再起動を実行する方法がある。しかし、この方法では再起動の期間中、サービスが停止することが問題となる。

【0004】

そこで、クラスタリング・ソフトウェアを使って物理的に複数のサーバでクラスタを構築し、WebサーバやDBMS (Database management System) といったサービスのフェールオーバーをする方法がある。

【0005】

これは、サービスを提供している現用系サーバとサービスを提供しない待機系サーバでクラスタを構成し、クラスタ内のサーバ間でハートビートと呼ばれるメッセージを相互に通信し合ったり、共有ディスクに定期的にタイムスタンプを書き込んだりして互いに障害が発生していないかチェックする。ハートビートが途切れたり、共有ディスクのタイムスタンプが適切に更新されていない場合は障害が発生したと検出し、障害があった現用系サーバで実行していたサービスを障害の起きていない待機系サーバで起動する (フェールオーバー) という動作をする。

【0006】

この物理サーバでクラスタを構築する方法では、(1)物理的に計算機を複数用意する必要があること、(2)ハートビート専用のネットワークを構築するために、ハートビートのためのルータ装置やネットワークインタフェースカードを追加する必要があること、(3)複数のサーバで同一のサービスを実行するために共通のデータディスクを持つ必要がある。

【0007】

対処する障害をソフトウェアに限定している場合、仮想計算機を用いることでこれらの課題を解決することができ、上記(1)に対しては、特開平9-2885

90号公報に示されているように、仮想計算機だけでクラスタを構成することで対処するものが知られている。これは、一つの物理計算機上で仮想計算機を複数稼動することで、OSやアプリケーション・ソフトウェアの多重化が可能となり、ソフトウェア障害に対して対処可能となる。

【0008】

上記(2)に対しては、特開平11-85547号公報に示されているように、仮想計算機間の通信を主記憶を使用したプロセス間通信で実現するものが知られており、仮想計算機間の通信用にルータやネットワークカード等のハードウェアを使用せずに仮想計算機のクラスタを構成することができる。

【0009】

上記(3)に対しては、クラスタ化した計算機間でデータディスクの共有化をするため、各サーバとディスクを接続するためにSCSI等のインタフェースを複数持ったディスク装置（マルチポートディスク、マルチポートRAIDなど）を使用するものが知られている。

【0010】

【特許文献1】

特開平9-288590号公報

【特許文献2】

特開平11-85547号公報

【非特許文献1】

ジム・グレイ、アンドレアス・ロイター 著／喜連川 優 監訳著、「トランザクション処理 概念と技法」、日経BP社発行、2001年10月29日発行、第122頁～第123頁

【0011】

【発明が解決しようとする課題】

しかしながら、上記従来例のように、複数の仮想計算機でマルチポートのディスク装置を共有する場合には、マルチポートのディスク装置（またはI/Oデバイス）が高価であるため、システムの製造コストが増大するという問題がある。

【0012】

また、上記従来例では、仮想計算機の現用系と待機系が共にマルチポートのディスク装置に対してアクセス可能となっているため、待機系のソフトウェアに障害が生じた場合でもディスク装置へアクセスが可能であるため、障害生じた待機系の不要なアクセスにより、現用系のディスクアクセスに悪影響を与えてしまう、という問題があった。

【0013】

そこで、本発明の課題は、安価なシングルポートのI/Oデバイスを採用しながらもフェールオーバー動作を可能にして、信頼性の向上と製造コストの低減を両立させることを目的とする。

【0014】

【課題を解決するための手段】

本発明は、計算機の制御プログラム上で構築された複数の仮想計算機と、前記計算機のPCIバスに接続されるとともに、前記複数の仮想計算機で共有されるI/Oデバイスと、を備えた計算機システムにおいて、

前記I/Oデバイスに配設されてPCIバスに接続された単一のポートと、前記複数の仮想計算機のうちのひとつと、前記ポートとの論理的な接続状態を設定するPCI接続割り当て手段と、前記仮想計算機からの制御信号に基づいて、前記PCI接続割り当て手段に設定された接続状態を更新するI/Oデバイス切り換え手段とを備え、前記仮想計算機は、PCI接続割り当て手段の設定に基づいてI/Oデバイスの変更を行う。

【0015】

特に、前記I/Oデバイス切り換え手段は、前記PCI接続割り当て手段の設定を更新するとともに、仮想計算機にI/Oデバイスの変更を通知する割り込みをかける割り込み手段と、を有し、前記割り込みを受けた仮想計算機は、前記PCI接続割り当て手段の設定に基づいてI/Oデバイスの変更を行う。

【0016】

【発明の効果】

したがって、仮想計算機から所定の制御信号が送られると、PCI接続されたI/Oデバイスの論理的（仮想的）な接続を変更するとともに、仮想計算機には

割り込みが通知されるので、この割り込みを受けた仮想計算機は、単一のポートの I/O デバイスをホットプラグすることで接続状態を切り換えることが可能となり、特に、現用系と待機系からなる複数の仮想計算機でこの I/O デバイスを共有する場合には、現用系に障害が生じると、I/O デバイスの単一のポートを待機系に切り換えるとともに、割り込み信号に基づいて待機系を現用系として起動させることができ、安価な単一のポートの I/O デバイスによって製造コストを抑制しながら信頼性を確保することが可能となる。

【0017】

【発明の実施の形態】

以下、本発明の一実施形態を添付図面に基づいて説明する。

【0018】

図1は、本発明の一実施形態を適用する計算機システムの構成図である。

【0019】

本実施形態の計算機システム200は、CPU201-0、201-1、201-2、201-3とCPUバス202とメモリコントローラ203とメモリバス204と主記憶205とI/Oバス216とI/Oブリッジ209とPCIバス210とPCIスロット212-0、212-1とPCIカード（PCIデバイス）111に接続されたディスク装置112から構成される。ただし、本発明を適用できる計算機システムは、図1で示されているCPUやI/Oブリッジ、PCIバス、PCIスロット、ディスク装置の数に限定されない。また、PCIカード111は、PCIスロット212-1に接続され、単一のポート（シングルポート）を有するものである。

【0020】

なお、PCIバス210、PCIカード111は、ACPI（Advanced Configuration and Power Interface Specification）2.0に準拠しホットプラグに対応したものである。ACPI 2.0の詳細については、<http://www.acpi.info/DOWNLOADS/ACPIspec-2-0b.pdf>に記載されるものである。

【0021】

複数の仮想計算機（以下、LPARとする）を構築するための制御プログラム

107と、この制御プログラム107によって構築されたLPAR上で動作するゲストOS206-0、…、206-nは主記憶上に置かれている。LPAR構築時に管理者によって各LPARごとにPCIスロットは割り当てられる。そのLPARとPCIスロットの割り当てを記述したPCIスロット割当表は制御プログラム107が保持している。なお、ゲストOS206-0、…、206-nは、ACPIのホットプラグ（以下、ホット・アッド）に対応しているものである。

【0022】

また、メモリコントローラ上には、PCIバス210に接続したPCIデバイス（I/Oデバイス）から主記憶205へのデータの読み書きをする要求（インバウンドアクセス）のアドレス変換を行うためのゲートキーパ110がある。

【0023】

本発明を適用した計算機システムの具体的な動作を図2を用いて説明する。

【0024】

図2の計算機100は、図1の計算機システムの動作を示すための概念的な構成図である。計算機100は、計算機のハードウェア109上で制御プログラム107が動作している。制御プログラム107が仮想計算機101（LPAR0）と仮想計算機102（LPAR1）を構築し、これらの複数の仮想計算機上でゲストOSが稼動している。各ゲストOS上では、クラスタリングソフト103、104が稼動し、ハートビート・ネットワーク107を使用して定期的にお互いに信号（ハートビート）を送り、相互に正常に稼動していることを確認しあっている。

【0025】

各LPARにどのPCIスロットが割り当てられているかは、制御プログラム107上のPCIスロット割当表108に記述されている。図4にPCIスロット割当表108の例400を示す。図4では、PCIスロット0と1がLPAR0に割り当てられている。また、PCIスロット割当表には、イネーブルフラグが各PCIスロットに割り当てられている。このイネーブルフラグが1のときは該当するPCIスロットに接続しているPCIデバイス（図2のPCIカード1

11) へのアクセスは許可されるが、イネーブルフラグが0のときは該当するPCIスロットに接続しているPCIデバイスへのアクセスは許可されない。つまり、図4の400の割り当て表では仮想計算機LPAR0のアクセスが許可され、割り当て表が図中401の状態では、仮想計算機LPAR0のアクセスは拒否され、また、割り当て表が図中402の状態では、LPAR1のみがアクセス可能となる。

【0026】

制御プログラム107は、PCIスロット割当表108を参照し、アクセスの許可判定118をし、ゲストOSからディスク装置112へのアクセス（アウトバウンドアクセス）115または116をセクタ117で選択する。アクセスを許可しないLPARからのアクセスは一般保護例外をアクセス要求元のOSへ発行する（図示なし）。

【0027】

ディスク装置112からのDMAアクセスのようなインバウンドアクセスは、ゲートキーパ110でアドレス変換され、各LPARに割り当てられたメモリ空間にアクセスする。ゲートキーパへのデータの書き込みは制御プログラム107が実行する。

【0028】

ここでは、LPAR0の仮想アドレス空間700とLPAR1の仮想アドレス空間701は、物理アドレス空間702に図7のようにマッピングされていると仮定する。PCIバス0に接続しているディスク装置112は、LPAR0に割り当てられているとすると、ゲートキーパ110には、図8の（a）のデータがセットされている。同様にディスク装置112がLPAR1に割り当てられているとすると、ゲートキーパ110は、図8の（b）のデータがセットされる。

【0029】

ゲストOS上のクラスタリングソフト103、104のようなユーザアプリケーションは、制御プログラム107に対して制御コマンド（図3）を発行することができる（113）。制御プログラム107は、コマンド制御ルーチン124でコマンドの解釈、実行をする。ゲストOS上の図3にこの制御コマンドの一例

を示す。制御コマンドは、例えば、図3に示すようにコマンド部分と仮想計算機の番号を指定する。

【0030】

図3(a)に`deact`コマンドは、指定された仮想計算機`LPAR`(この図では、`LPAR0`)をディアクティベートするコマンドである。このコマンドを実行すると`LPAR0`は非活性化、つまり、物理的な計算機でいう電源オフの状態になる。

【0031】

同じく図3(b)の`act`コマンドは、指定された仮想計算機`LPAR`(この図では、`LPAR0`)をアクティベートするコマンドである。このコマンドを実行すると`LPAR0`は、活性化、つまり物理的な計算機でいう電源オンの状態になる。

【0032】

`add_pci`コマンド(図3(c))は指定された仮想計算機`LPAR`(ここでは、`LPAR0`)に割り当てられた`PCI`スロット(この`PCI`スロットにインストールしてある`PCI`デバイスも含めて)をコマンド発行元の`LPAR`に論理的に接続する。このコマンドの発行元`LPAR`がアクティブ状態であれば図5に示す`Hot Add`動作が実行される。`add_pci`コマンドの動作を図5を用いて説明する。

【0033】

図5は、あるゲスト`OSm`の稼動する`LPARm`に割り当てられている`PCI`スロット`s`をゲスト`OSn`の稼動する`LPARn`に論理的にホット・アッドする処理の一例を示すフローチャートで、制御プログラム107で行われるものである。図中、`m`、`n`、`s`は整数で $m \neq n$ である。

【0034】

ステップ500では、ゲスト`OSn`が`add_pci`コマンドを発行し、`PCI`スロット`s`を論理ホット・アッドを開始する。

【0035】

ステップ501では、`PCI`スロット`s`の状態を判定して、`PCI`スロット`s`

に P C I デバイスが挿されていない場合、ステップ 5 0 2 に進み、制御プログラムは P C I デバイスを挿すようにコマンド発生元に指示する。一方、P C I スロット s に P C I デバイスが挿入されていればステップ 5 0 3 に進む。

【0036】

ステップ 5 0 2 は、P C I デバイスが P C I スロット s に挿されていない場合であり、管理者が P C I デバイスを P C I スロット s に挿入するよう指示を行い、P C I デバイスが挿入されればステップ 5 0 3 に進む。

【0037】

ステップ 5 0 3 では、P C I デバイスが P C I スロット s に挿されている場合、制御プログラム 1 0 7 は O S n から P C I スロット s へのアクセスを許可する処理をする。具体的には、P C I スロット割当表 1 0 8 で P C I スロット s の接続先を L P A R n とし、イネーブルフラグを 1 に設定する。例えば、L P A R 0 に割り当てられている P C I スロット 0 の割当を L P A R 1 へ切り換えるには、図 4 の 4 0 0 の状態から 4 0 2 の状態にする更新を行う。

【0038】

また、ゲートキーパに登録されている仮想メモリ空間の物理メモリ空間のマッピング情報を L P A R n (ゲスト O S n) に対応するデータに書き換える。例えば、L P A R 0 と L P A R 1 のアドレス空間を図 7 のように設定していて、L P A R 0 に割り当てられているゲートキーパの設定を、図 8 の (a) の状態から (b) の状態にする更新を行う。

【0039】

ステップ 5 0 4 では、制御プログラム 1 0 7 は、論理的な S C I (System Call Interrupt) 割り込みをゲスト O S n に対して発行する。なお、S C I 割り込みは、A C P I で定義されているものである。

【0040】

ステップ 5 0 5 で、ゲスト O S n は、G P E (General-Purpose Event) レジスタの内容を読みに行く。なお、上記 S C I 割り込み、G P E レジスタについては、A C P I 2. 0 で規定されるものである。

【0041】

ステップ506では、制御プログラム107は、ゲストOS_nからのGPEレジスタアクセスをトラップし、自身にハードコーディングされているGPEレジスタの内容 (insertion event) をゲストOS_nへ返す。

【0042】

ステップ507では、ゲストOS_nのACPI処理ルーチンでPCIスロットsのPCIデバイスの設定を開始する。

【0043】

ステップ508では、ゲストOS_nまたはゲストOS_n上のアプリケーションは、追加したPCIデバイスを使用できるように処理をする。具体的には、PCIデバイスに接続されているディスク装置をマウント等をする。

【0044】

以上の処理により、PCIデバイスは、OS_mからOS_nに割り当てられ、再起動などを要することなくOS_nからこのPCIデバイスを制御することが可能となるのである。

【0045】

次に、図3(d)に示した、rem_pciコマンド(図3(d))は指定された仮想計算機LPAR_nに割り当てられたPCIスロット(このPCIスロットにインストールしてあるPCIデバイスも含めて)を論理的に取り外すHot Remove動作(図6参照)が実行される。

【0046】

図6は、あるゲストOS_mの稼動するLPAR_mからrem_pciコマンドが発行され、ゲストOS_nの稼動するLPAR_nに割り当てられているPCIスロットsをホット・リムーブする処理の一例を示すフローチャートで、制御プログラム107で行われるものである。なお、図中m、n、sは整数で、m≠nである。

【0047】

ステップ600では、ゲストOS_mから、rem_pciコマンドが発行され、ゲストOS_nの稼動するLPAR_nのPCIスロットsを論理ホット・リムーブを開始する。

【0048】

ステップ601では、制御プログラム107は、LPAR_nからPCIスロットsへのアクセスを不許可に変更する。具体的には、制御プログラム内のPCIスロット割当表108のPCIスロットsのイネーブルフラグを0にする。例えば、LPAR₀に割り当てられているPCIスロット0の割当を、図4の400の状態から401の状態にする更新を行う。

【0049】

ステップ602では、制御プログラム107は、PCIスロットsに接続していたLPAR_nがアクティブであるかどうか判定し、アクティブであればステップ603へ進む一方、アクティブでなければそのまま処理を終了する。

【0050】

ステップ603では、ステップ602でLPAR_nがアクティブの場合、制御プログラムが論理SCI割り込みをゲストOS_nに発行する。

【0051】

ステップ604では、ゲストOS_nは、GPEレジスタの内容を読みに行く。

【0052】

ステップ605では、制御プログラム107は、ゲストOS_nからのGPEレジスタアクセスをトラップし、自身にハードコーディングされているGPEレジスタの内容(eject request)をゲストOS_nへ返す。

【0053】

ステップ606では、ステップ602でLPAR_nがアクティブで無い場合、またはステップ605実行後、ゲストOS_nでPCIスロットsのPCIデバイスの使用を停止する。具体的には、PCIスロットsのPCIデバイスに接続しているディスク装置をアンマウントする。

【0054】

以上の処理により、PCIデバイスは、OS_nに対する割り当てが無くなって、再起動などを要することなくOS_nからこのPCIデバイスをアンマウントすることが可能となるのである。

【0055】

次に、本発明を適用した計算機システムがサービスのフェールオーバーする過程を図2と図9を用いて説明する。

【0056】

フェールオーバーをする際、障害が起きている仮想計算機L P A Rで、ゲストO Sそのものが障害を起こしているケースと、O Sに異常は無いがフェールオーバー対象のサービスのみが障害を起こしているケースが考えられる。

【0057】

図9では、前者の障害に対処するケース、つまり障害側L P A RのO SをシャットダウンせずにL P A Rのデアクティベートするケース(a)と、後者の障害に対処するケース、つまり障害側L P A RのO SのシャットダウンをしてからL P A Rのデアクティベートをするケース(b)がある。まず、(a)のケースについて説明した後、(b)については(a)との差分のみ説明する。(a)、(b)どちらの動作をするかは、クラスタリングソフトの仕様、あるいは設定で決まる。

【0058】

この実施形態では説明のため以下の動作状態になっていると仮定する。

【0059】

図2のL P A R 0がサービスを提供する現用系でL A P R 1が待機系となっている。また、ディスク装置112は、現用系L P A Rで運用するサービスが使用しているとする。例えば、ディスク装置112にはデータベースのテーブルが格納されているとする。また、ディスク装置112はP C Iスロット0に接続されているP C Iカード(P C Iデバイス)111に接続し(123)ている。

【0060】

制御プログラム107のP C Iスロット割当表108は、図4の400の状態になっている。また、ゲートキーパ110は、図8の状態になっているものとする。

【0061】

<図9のケース(a)>

まず、ステップ1100では、L P A R 1のクラスタリングソフト104がL

PAR0で障害が発生していることを検出する。

【0062】

ステップ1101では、クラスタリングソフト104から制御プログラム107に対して障害発生側のLPAR0をデアクティベートするdeactコマンド(図3の(a))を発行する(113)。制御プログラム107は、このdeactコマンドをコマンド制御ルーチン124で解読し、LPAR0をデアクティベートする(114)。

【0063】

ステップ1102では、クラスタリングソフト104から制御プログラム107に対して、障害発生側LPAR0のPCIデバイスをLPAR1にホット・リムーブするためにrem_pciコマンド(図3の(d))を発行する(113)。rem_pciコマンドの動作は、上述の図5のホット・リムーブルーチンの動作説明で、m=0、n=1と読み替えた動作をする。

【0064】

ステップ1103では、クラスタリングソフト104から制御プログラム107に対して、障害発生側LPAR0のPCIデバイスをLPAR1にホット・アッドするためにadd_pciコマンド(図3の(c))を発行する(113)。add_pciコマンドの動作は上述の図5のホット・アッドルーチンの動作説明で、m=1、n=0と読み替えた動作をする。

【0065】

ステップ1104では、クラスタリングソフト104が、add_pciコマンドで接続したディスク112へアクセスできることを確認した後、障害側であるLPAR0で実行していたサービスをLPAR1のゲストOS1上で起動する。

【0066】

ステップ1105では、クラスタリングソフト104から制御プログラム107に対して障害発生側のLPAR0をアクティベートするactコマンドを発行(113)し、待機系として再起動する。制御プログラム107は、このactコマンドをコマンド制御ルーチン124で解読し、LPAR0をアクティベート

する(114)。

【0067】

以上の処理は、図10に示すように、障害が発生したL P A R 0は、デアクティブートされた後に、P C I デバイスはL P A R 0からホット・リムーブされ、その後、L P A R 1へホット・アッドされて、P C I スロットの割り当て表108が更新される。この後、L P A R 1ではL P A R 0で行われていたサービスが開始されて現用系(図中、稼動系)としてクライアントの要求を受けると共に、障害が発生したL P A R 0は待機系として再起動され、L P A R 0、1は現用系と待機系が切り替わる。

【0068】

これにより、単一のポートを備えたP C I デバイスは、現用系のL P A R からホット・リムーブされた後に、待機系へホット・アッドされてP C I デバイスの論理的な接続が変更されるので、必ず単一のO S の制御下に置かれながらも仮想的(論理的)にリムーブ(接続終了)とアッド(接続開始)が行うことができるので、再起動などの必要がなくなって、信頼性の高い仮想計算機システムをシングルポートのP C I デバイスを用いることで安価にて提供することが可能となり、信頼性の向上と製造コストの低減を両立させることが可能となるのである。

【0069】

<図9のケース(b)>

ステップ1106は、上記ステップ1100と同一の動作。

【0070】

ステップ1107は、ステップ1102と同一の動作。

【0071】

ステップ1108は、クラスタリングソフト104からL P A R 0のゲストO S 0に対して稼動中のサービスの停止とO S のシャットダウンを指示する(107)。クラスタリングソフト103は稼動中のサービスの停止とO S のシャットダウンを実行する。

【0072】

ステップ1109は、ステップ1101と同一の動作。

【0073】

後の動作は、上記で既に説明済みである。

【0074】

この場合では、図11で示すように、障害が生じたL P A R 0からP C Iデバイスがホット・リムーブされた後に、L P A R 0のデアクティベートが行われ、その後は、上記ステップ1103以降と同様にL P A R 1にP C Iデバイスがホット・アッドされてP C Iスロットの割り当て表108が更新され、さらにL P A R 1でサービスが開始されて現用系と待機系が切り替わってから、障害の生じたL P A R 0が待機系として再起動されるのであり、単一のポートを備えたP C Iデバイスによって現用系と待機系を切り換えることが可能となる。

【0075】

従来の物理計算機におけるP C Iデバイスのホットプラグは、P C Iデバイスを抜き差ししたことでハードウェアがG P E (General Purpose Event) を起こすためのS C I (System Control Interrupt) 割り込みをO Sに送ることでホットプラグ処理が始まる。S C I割り込みを受けたO SはG P Eレジスタの内容を読みに行ってP C Iデバイスのホットプラグがあったことを知り、後はA C P Iの規約に記載されたとおりのホットプラグ処理を開始する。

【0076】

一方、本発明の特徴は、このS C I割り込み発生部分とG P Eレジスタを仮想計算機の制御プログラム107でエミュレートすることで論理的なP C Iデバイスのホット・プラグを実現することにある。

【0077】

すなわち、制御プログラム107の管理下で複数の仮想計算機L P A Rが動作する仮想計算機システムで、現用系仮想計算機L P A R 0と待機系仮想計算機L P A R 1がある場合、待機系仮想計算機L P A R 1が現用系仮想計算機L P A R 0の障害を検出し、制御プログラム107に報告すると、制御プログラム107が、待機系計算機L P A R 1にS C I割り込みを仮想的に発行する。そして、O SからのG P Eレジスタの参照リクエストに対して制御プログラム107がホットプラグ・イベントが起きたことを知らせるデータをハードウェアの代わりに返

送する。このとき、制御プログラム107は、P C I デバイスを接続する仮想計算機L P A R 1に対してアクセスを許可するように各仮想計算機をアクセスを許可するアドレスの設定を変更する。

【0078】

こうして、S C I 割り込みを受けた仮想計算機L P A R 1のO SのA C P I 処理ルーチンがP C I デバイスのホットプラグ処理を行うことで論理的なP C I デバイスのホットプラグが可能となるのであり、上記仮想計算機L P A R 1にP C I デバイスが新たに追加されることになる。

【0079】

また、S C I 割り込みを受けた仮想計算機L P A R 0のO SのA C P I 処理ルーチンがP C I デバイスのホットリムーブ処理を行うことで論理的なP C I デバイスのホットリムーブが可能となる。

【0080】

これらの論理的（仮想的）なホット・プラグ。ホット・リムーブにより、単一のポートを備えたP C I デバイス（P C I カード111）は、仮想計算機L P A R 0から切り離される一方、仮想計算機L P A R 1に接続され、現用系と待機系の仮想計算機で共有される。また、P C I デバイスは、現用系に常に接続されるので、待機系はP C I デバイスにアクセスすることができず、このため、待機系に障害が発生しても、P C I デバイスのアクセスを防ぐことができ、より信頼性の高い計算機システムを構築することが可能となる。

【0081】

なお、上記第一の実施形態ではP C I カードに接続する装置をディスク装置の場合のみ説明したが、ネットワークカード等のディスク以外の装置に対して適用することができる。

【0082】

また、現用系と待機系の仮想計算機を切り換えたときには、現用系のC P Uの割当比率が高くなるようにしても良く、例えば、現用系には90%、待機系には10%のC P Uの割当比率（割当時間）に設定することにより、多重化に伴う性能低下を抑制できる。このC P Uの割当比率の変更は、制御プログラム107に

おいて行えばよい。

【0083】

次に、図12は、本発明の第2実施形態を適用する計算機システムの構成図である。

【0084】

上記第1実施形態の仮想計算機に変わって、計算機ハードウェアを物理的に分割し、一台の計算機を複数の計算機にする物理分割計算機（PPAR）に本発明を適用した実施形態を示す。

【0085】

本実施形態の計算機システム900は、CPU903-0、903-1、903-2、903-3、906-0、906-1、906-2、906-3とCPUバス907、908とメモリコントローラ909、910とスイッチ904、905とメモリコントローラ間ネットワーク911とメモリバス912、913と主記憶916、917とI/Oバス914、915とI/Oブリッジ930、931とPCIバス922、923とPCIスロット924-0、924-1、925-0、925-1とPCIカード926、927とディスク装置929、936、SVP（Service Procesor）941、コントロールバス940、コンソール942から構成される。

【0086】

ただし、本発明を適用できる計算機システムは、図12で示されているCPUやI/Oブリッジ、PCIバス、PCIスロット、ディスク装置の数に限定されない。

【0087】

計算機システム900は、物理分割計算機901（以下、PPAR0）と物理分割計算機902（以下、PPAR1）に物理的に分割できる。この分割は、管理者がコンソール942から設定し、SVP941がメモリコントローラ内909、910内のスイッチ904、905を切り替えてメモリコントローラ間ネットワークを無効にすることで実現される。

【0088】

管理者はコンソール942からSVP941上のPCIスロット割り当て表950を設定することができる。PCIスロット割り当て表950の内容は、SVP941がコントロールバス940、メモリコントローラ909、910、メモリバス912、913を通して、主記憶916、917上の制御プログラム920、921に反映される。OS918、919からは制御プログラム920、921で指定されたPCIスロットが割り当てられることになる。

【0089】

OS918、919からは主記憶916、917にある制御プログラム920、921が指定したPCIスロット924-0、1、925-0、1に接続されているPCIデバイス926または927しか接続されていないように見える。このため、メモリコントローラ909とI/Oブリッジ931の間には、内部バス1943が設けられ、メモリコントローラ910とI/Oブリッジ930の間には、内部バス1942が設けられる。なお、この制御プログラム920、921は一般にBIOS (Basic Input Output System) と同一の機能を有するファームウェアである。

【0090】

この第2の実施形態では、図12のPPAR0とPPAR1でクラスタを構成し、例えば、PPAR0が現用系、PPAR1が待機系に割り当てられている。以下では、本発明をPPARでクラスタを構成した場合のフェールオーバー動作をするにあたり、第1の実施形態と異なる点のみ説明することにする。

【0091】

上記第1実施形態の図3、図4でLPARと記述されている箇所がすべてPPARとなり、図2のPCIスロットの割り当て表108と図12のPCIスロットの割り当て表950は等価となる。

【0092】

同じく、図3で示されているコマンドは主記憶上の制御プログラムからコントロールバス940を介してSVPへ送られる。各PPARのデアクティベートやアクティベートはSVP941で実行される。つまり、SVP941によりコマンドが対象としているPPARの起動やシャットダウンの制御が実行される。

【0093】

図5のホット・アッド処理では、第一の実施形態と異なる箇所を列挙する。

【0094】

上記図5の全ステップについて、LPARをPPARに置き換える。

【0095】

同じく図5のステップ503は、ゲートキーパの設定変更がなくなり、OS上の制御プログラムからSVP941に対してPCIスロット割り当て表950の変更をする要求をコントロールバス940経由で送る。SVP941は、PCIスロット割り当て表850を更新し、PCIスロット割り当てをPPAR_nに変更する。

【0096】

図6のホット・リムーブ処理で、第一の実施形態と異なる箇所を列挙する。

【0097】

全ステップについて、LPARをPPARに置き換える。
ステップ601で、制御プログラムは、OS_nから対象PCIスロットへのアクセスを不許可に変更するためにSVP941に対して、PCIスロット割り当て変更の要求をコントロールバス940経由で送る。SVP941はPCIスロット割り当て表950でPCIスロット_sのイネーブルフラグを0に設定する。

【0098】

図9のフェールオーバ処理において、第一の実施形態と異なる箇所を列挙する。
。

【0099】

全ステップについて、LPARと記述された部分がすべてPPARに置き換わる。

【0100】

以上の変更により第2の実施形態では、SVP941のPCIスロットの割り当て表950を更新することにより、PPAR₀とPPAR₁で共有するシングルポートのPCIデバイス（カード）926、927を仮想的にホット・リムーブ、ホット・アッドを行って、フェールオーバ処理においてPCIデバイスをP

P A R 間で論理的に付け替えが実現できるのである。

【0101】

なお、図12において、データディスクとしてディスク装置929のみを使用している場合で、現用系をP P A R 0からP P A R 1へ切り換える際には、P C Iカード926についてP C Iスロット割り当て表950を更新し、現用系に切り替わったP P A R 1はメモリコントローラ905、内部バス1942を介してディスク装置929にアクセスすることができる。

【0102】

図13～図17は、第3の実施形態を示す。

【0103】

第3の実施形態では、図13で示すように、P C Iカード1002に搭載されているROM1003上にホット・プラグ処理を起動するための割り込み信号を論理的に発行する制御プログラムが搭載されているケースである。

【0104】

図13では、P C Iバス1000にP C Iスロット1001があり、このP C Iスロット1001にP C Iカード1002が接続されている。P C Iカード1002は、信号線1004を介してディスク装置1005と接続されている。P C Iカード1002には、制御プログラムを格納しているROM1003が搭載されている。

【0105】

本実施形態は第一の実施形態の変形例であり、上記図2で示された仮想計算機システムでのP C Iカード111が、図13のP C Iカード1002と等価であると考えてよい。

【0106】

第一の実施形態と異なる点は、上記図5のホット・アッド処理において、ステップ504で第一の実施形態ではS C I割り込み信号を主記憶上の制御プログラム107が発行しているが、本実施形態ではROM1003上に格納されている制御プログラムから発行することが異なり、その他の点は上記図5と同様である。

。

【0107】

具体的には、図14のタイムチャートで示すように、図2の主記憶上の制御プログラム107がLPAR（図中ゲストOS）から受け取ったadd_pciコマンドをコマンド制御ルーチン124が解読し、ハードウェア109上のPCIカード111にSCI割り込み信号を送るように要求する。SCI割り込み要求を受けたPCIカード1002は、図13のROM1003上の制御プログラムがSCI割り込みを発行する。この割り込み信号を制御プログラム107は、add_pciコマンド発行元のLPARへ送ることである。これにより、LPARはGPEレジスタを参照した後、ACPI処理ルーチンを実行して、PCIカード1002のマウントを行う。

【0108】

同様に図6のホット・リムーブ処理で第一の実施形態と異なる点は、ステップ603で制御プログラム107がSCI割り込みを発行していたことを、図13のROM1003上に格納されている制御プログラムから発行することであり、その他は上記第1実施形態と同様である。

【0109】

具体的には、図15のタイムチャートで示すように、図2の制御プログラム107がLPAR_mから受け取ったrem_pciコマンドをコマンド制御ルーチン124が解読し、ハードウェア109上のPCIカード1002にSCI割り込み信号を送るように要求する。SCI割り込み要求を受けたPCIカード1002は、図13のROM1003上の制御プログラムがSCI割り込みを発行する。この割り込み信号を主記憶上の制御プログラム107は、rem_pciコマンドで指定したLPAR_nへ送る。この後、LPAR_nは、GPEレジスタを参照してからACPI処理ルーチンを実行する。

【0110】

以上により、PCIカード1002のROM1003に、ホット・プラグ処理を起動するための割り込み信号を論理的に発行する制御プログラムを格納することで、前記第1実施形態と同様に、PCIカード1002の仮想的なホット・アッド及びホット・リムーブを実現することができる。これにより、単一のポート



を備えた P C I デバイスを、複数の仮想計算機の間で仮想的に切り換えて、システムの信頼性を向上できるのである。

【0111】

次に、図 16 は図 13 の構成を上記第 2 実施形態に示したように、物理的な計算機へ適用した場合のホット・アッドまたはホット・リムーブのタイムチャートの一例を示し、図 12 の P C I カード 926、928 及びディスク装置 929、936 を図 13 に置き換えたものである。

【0112】

ホット・アッドの処理は上記図 14 と同様であり、割り込み信号の送り先が物理計算機（図 12 の P P A R 0、1 及び O S）である点が異なるだけである。

【0113】

上記図 5 のホット・アッド処理と同様に、ステップ 504 で第一の実施形態では S C I 割り込み信号を主記憶上の制御プログラム 107 が発行しているが、本実施形態では R O M 1003 上に格納されている制御プログラムから物理計算機の O S に発行する点が異なり、その他は前記第 1 実施形態と同様である。

【0114】

図 16 のタイムチャートで示すように、図 12 の主記憶上の制御プログラム 920 または 921 が O S から受け取った a d d _ p c i コマンドをコマンド制御ルーチンが解釈し、P C I カード 1002 に S C I 割り込み信号を送るように要求する。S C I 割り込み要求を受けた P C I カード 1002 は、図 13 の R O M 1003 上の制御プログラムが S C I 割り込みを発行する。制御プログラム 920、921 は、a d d _ p c i コマンド発行元の O S へ送る。これにより、O S は G P E レジスタを参照した後、A C P I 処理ルーチンを実行して、P C I カード 1002 のマウントを行う。

【0115】

同様に図 6 のホット・リムーブ処理で第一の実施形態と異なる点は、ステップ 603 で制御プログラム 107 が S C I 割り込みを発行していたことを、図 13 の R O M 1003 上に格納されている制御プログラムから発行することである。

【0116】

すなわち、図16のタイムチャートで示すように、図12主記憶上の制御プログラム920または921がOSから受け取ったrem_pciコマンドをコマンド制御ルーチンが解釈し、PCIカード1002にSCI割り込み信号を送るように要求する。SCI割り込み要求を受けたPCIカード1002は、図13のROM1003上の制御プログラムがSCI割り込みを発行する。この割り込み信号を主記憶上の制御プログラム920または921は、rem_pciコマンドで指定したOSへ送る。その後、OSは、GPEレジスタを参照してからACPI処理ルーチンを実行する。

【0117】

以上により、PCIカード1002のROM1003に、ホット・プラグ処理を起動するための割り込み信号を論理的に発行する制御プログラムを格納することで、図14、図15に示した仮想計算機の場合と同様に、物理計算機においても、PCIカード1002の仮想的なホット・アッド及びホット・リムーブを実現することができる。

【0118】

次に、図17は、図13に示したディスク装置1005に障害が発生した場合のタイムチャートを示し、計算機側は、図16と同様の物理計算機とする。

【0119】

SCSI等のインターフェースを備えたPCIカード1002は、ROM1003上の制御プログラムによってディスク装置1005に回復不能な障害が発生したことを検知すると、物理計算機のGPEレジスタを設定するとともに、OSに対してSCI割り込みを発生する。

【0120】

この割り込み信号を受信したOSは、GPEレジスタを参照してからACPI処理ルーチンを実行し、障害の発生したPCIカード1002をホット・リムーブする。

【0121】

こうして、PCIカード1002に接続されたデバイスに障害が生じた場合にも、GPEレジスタを設定してOSへ割り込みをかけることにより、PCIデバ

イスのホット・リムーブを行うことができるのである。

【図面の簡単な説明】

【図 1】

本発明の仮想計算機システムの一実施形態を示すブロック図である。

【図 2】

同じく、仮想計算機システムの一例で、各構成要素間のインタラクションを示すブロック図である。

【図 3】

仮想計算機、あるいは物理分割計算機の制御プログラムで実行するコマンドの形式を示す図である。

【図 4】

P C I スロットの割り当てや有効／無効状態を管理するための P C I スロット割当表である。

【図 5】

ホット・アッド処理のフローチャートである。

【図 6】

ホット・リムーブ処理のフローチャートである。

【図 7】

仮想計算機システムにおける仮想計算機ごとに割り当てられた仮想アドレス空間と物理メモリの配置関係を示す図である。

【図 8】

ゲートキーパが所有する P C I バスと、その P C I バスからアクセスする物理アドレスを割り出す表を示す図である。

【図 9】

フェールオーバー動作のフローチャートで、(a) は、障害側 L P A R の O S をシャットダウンせずに L P A R のデアクティベートするケースで、(b) は障害側 L P A R の O S のシャットダウンをしてから L P A R のデアクティベートをするケースを示す。

【図 10】

図9の(a)のケースに対応するタイムチャート。

【図11】

図9の(b)のケースに対応するタイムチャート。

【図12】

第2の実施形態を示し、物理分割計算機の一例を示すブロック図である。

【図13】

第3の実施形態を示し、PCIカードの一例を示すブロック図である。

【図14】

仮想計算機でホット・アッド処理を行う場合の一例を示すタイムチャート。

【図15】

仮想計算機でホット・リムーブ処理を行う場合の一例を示すタイムチャート。

【図16】

物理計算機でホット・アッド（またはホット・リムーブ）処理を行う場合の一例を示すタイムチャート。

【図17】

同じく、PCIデバイス側の障害発生によるホット・リムーブの一例を示すタイムチャート。

【符号の説明】

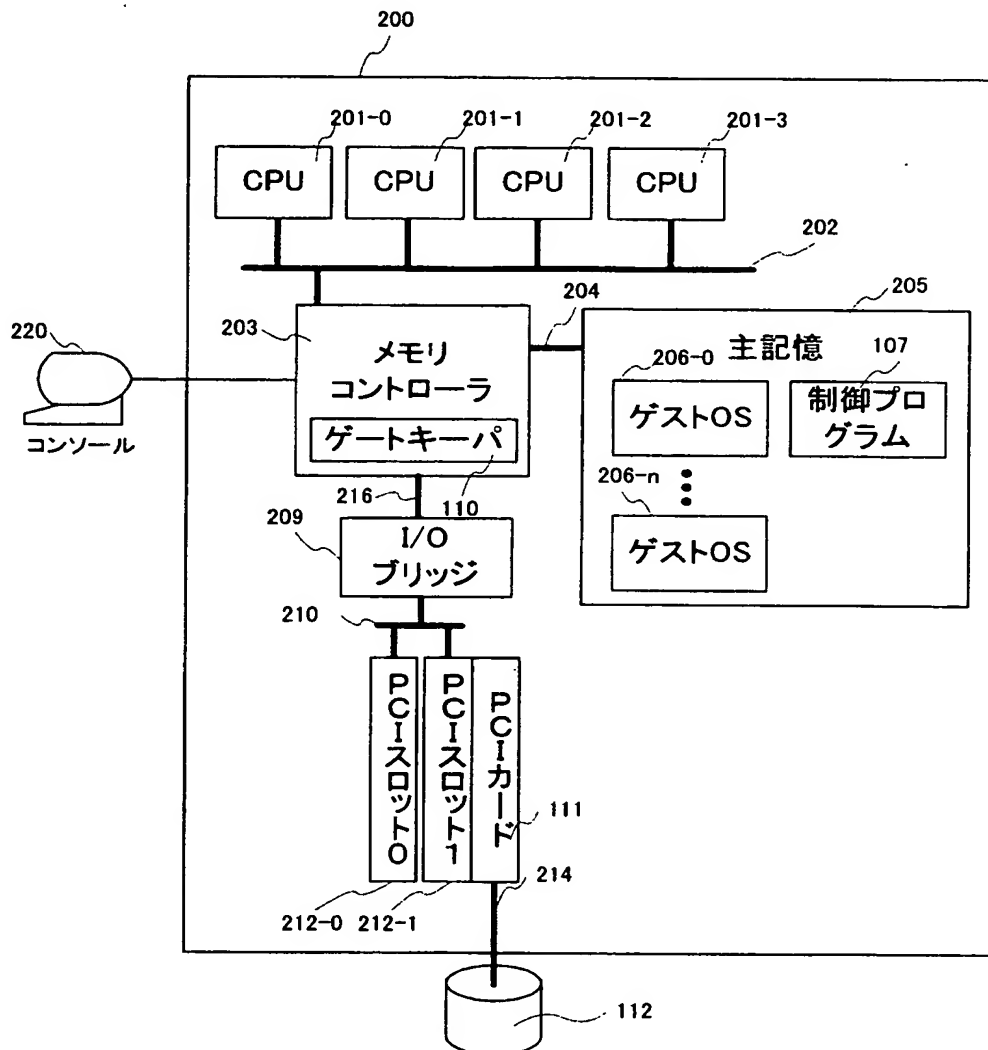
- 100 物理計算機
- 101 LPAR0
- 102 LPAR1
- 103、104 クラスタリングソフト
- 105、106 ACPI処理ルーチン
- 107 制御プログラム
- 124 コマンド制御ルーチン
- 108 PCIスロット割当表
- 110 ゲートキーパ
- 111 PCIカード
- 112 ディスク装置

203 メモリコントローラ
205 主記憶
206、206-n ゲストOS
209 I/Oブリッジ
212-0 PCIスロット0
212-1 PCIスロット1
220 コンソール装置
400、401、402 PCIスロット割当表
700 LAPR0の仮想アドレス空間
701 LAPR1の仮想アドレス空間
702 物理アドレス空間
904、905 スイッチ
920、921 制御プログラム
940 コントロールバス
941 SVP
942 コンソール
950 PCIスロット割当表
1000 PCIバス
1001 PCIスロット
1002 PCIカード
1003 ROM
1005 ディスク装置

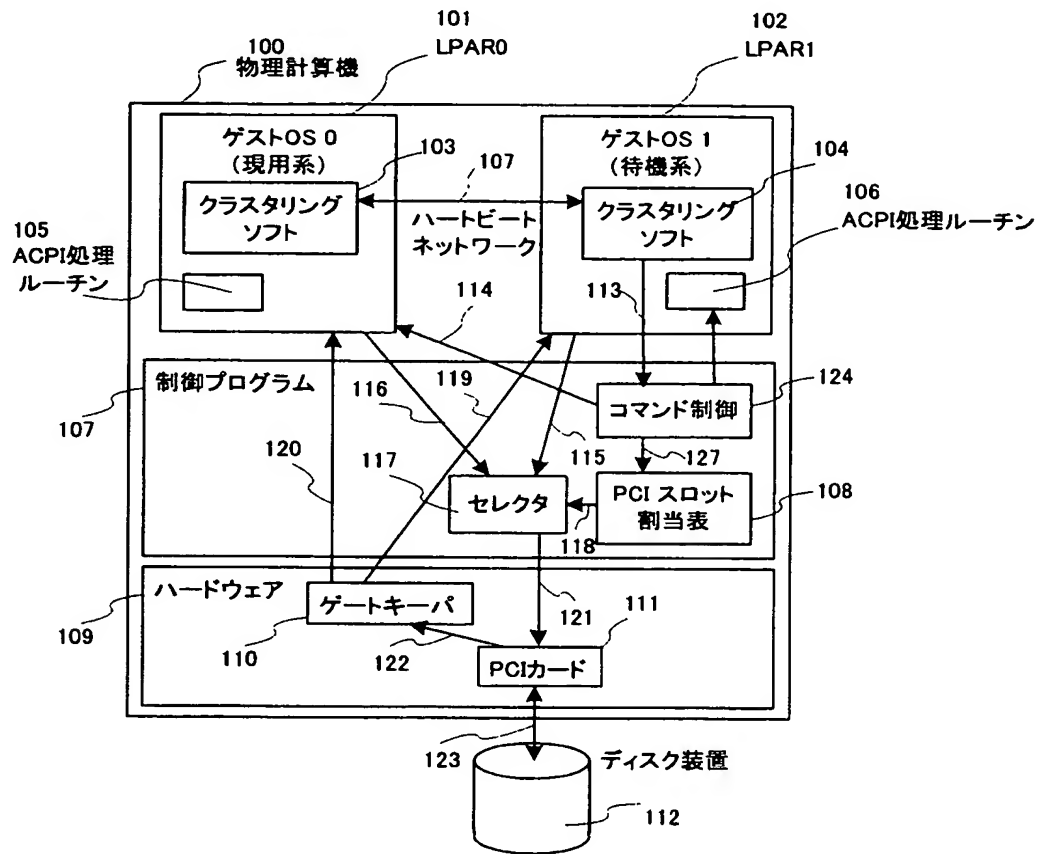
【書類名】

図面

【図 1】



【図 2】



【図 3】

コマンド名 仮想計算機番号			
(a)	<table border="1"> <tr> <td>deact</td><td>LPAR = 0</td></tr> </table>	deact	LPAR = 0
deact	LPAR = 0		
(b)	<table border="1"> <tr> <td>act</td><td>LPAR = 0</td></tr> </table>	act	LPAR = 0
act	LPAR = 0		
(c)	<table border="1"> <tr> <td>add_pci</td><td>LPAR = 0</td></tr> </table>	add_pci	LPAR = 0
add_pci	LPAR = 0		
(d)	<table border="1"> <tr> <td>rem_pci</td><td>LPAR = 0</td></tr> </table>	rem_pci	LPAR = 0
rem_pci	LPAR = 0		

【図 4】

PCI slot #	接続先	イネーブル フラグ
0	LPAR0	1
1	LPAR0	1

400

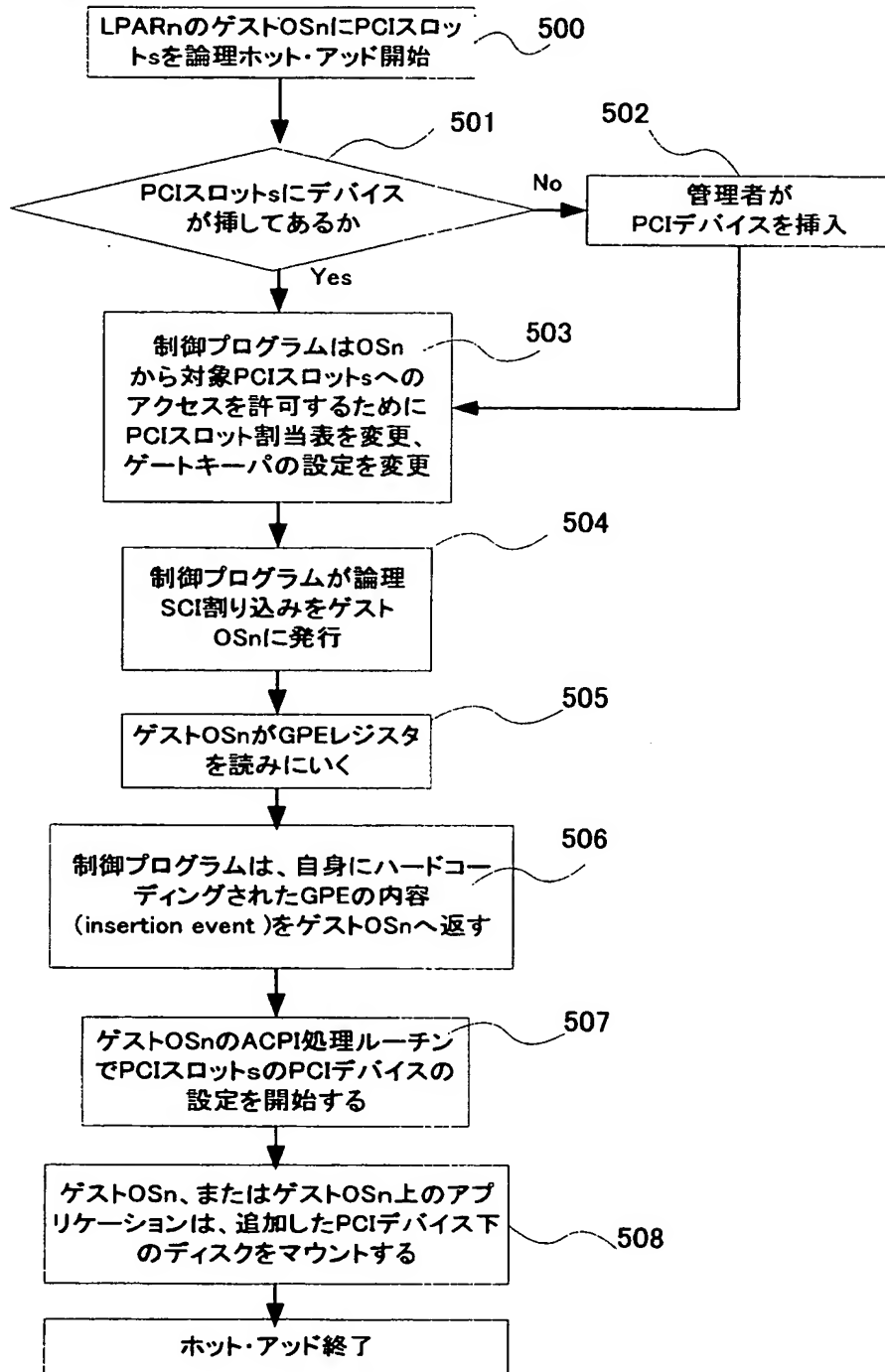
PCI slot #	接続先	イネーブル フラグ
0	LPAR0	0
1	LPAR0	0

401

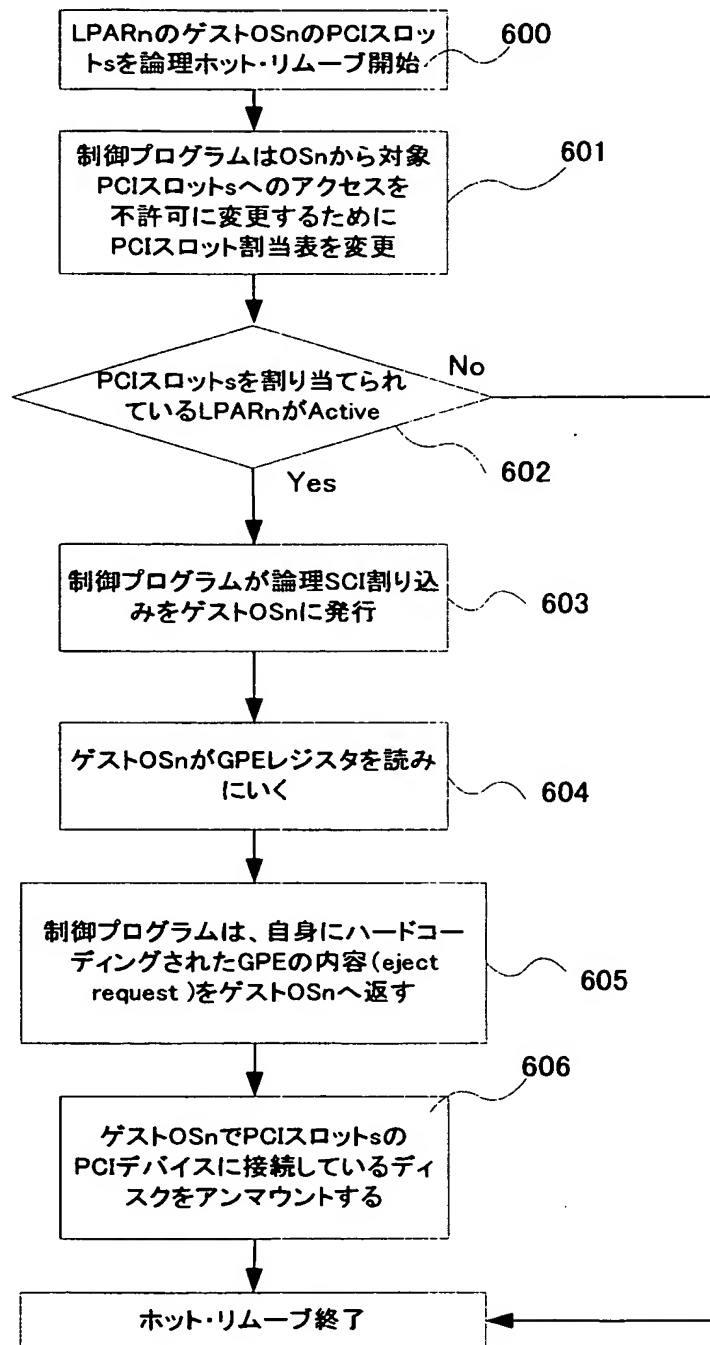
PCI slot #	接続先	イネーブル フラグ
0	LPAR1	1
1	LPAR1	1

402

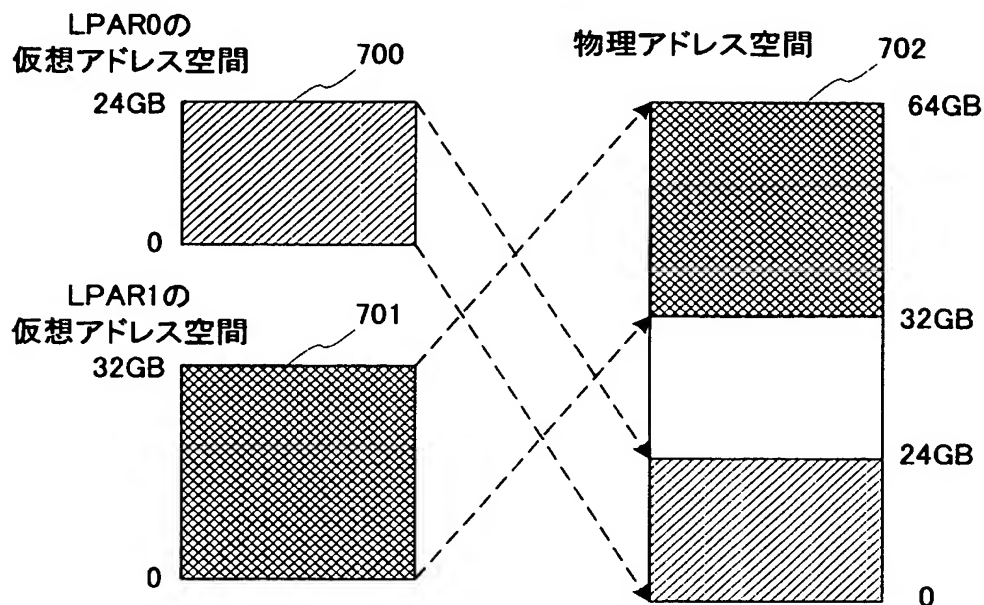
【図 5】

ホット・アッド処理

【図 6】

ホット・リムーブ処理

【図 7】



【図 8】

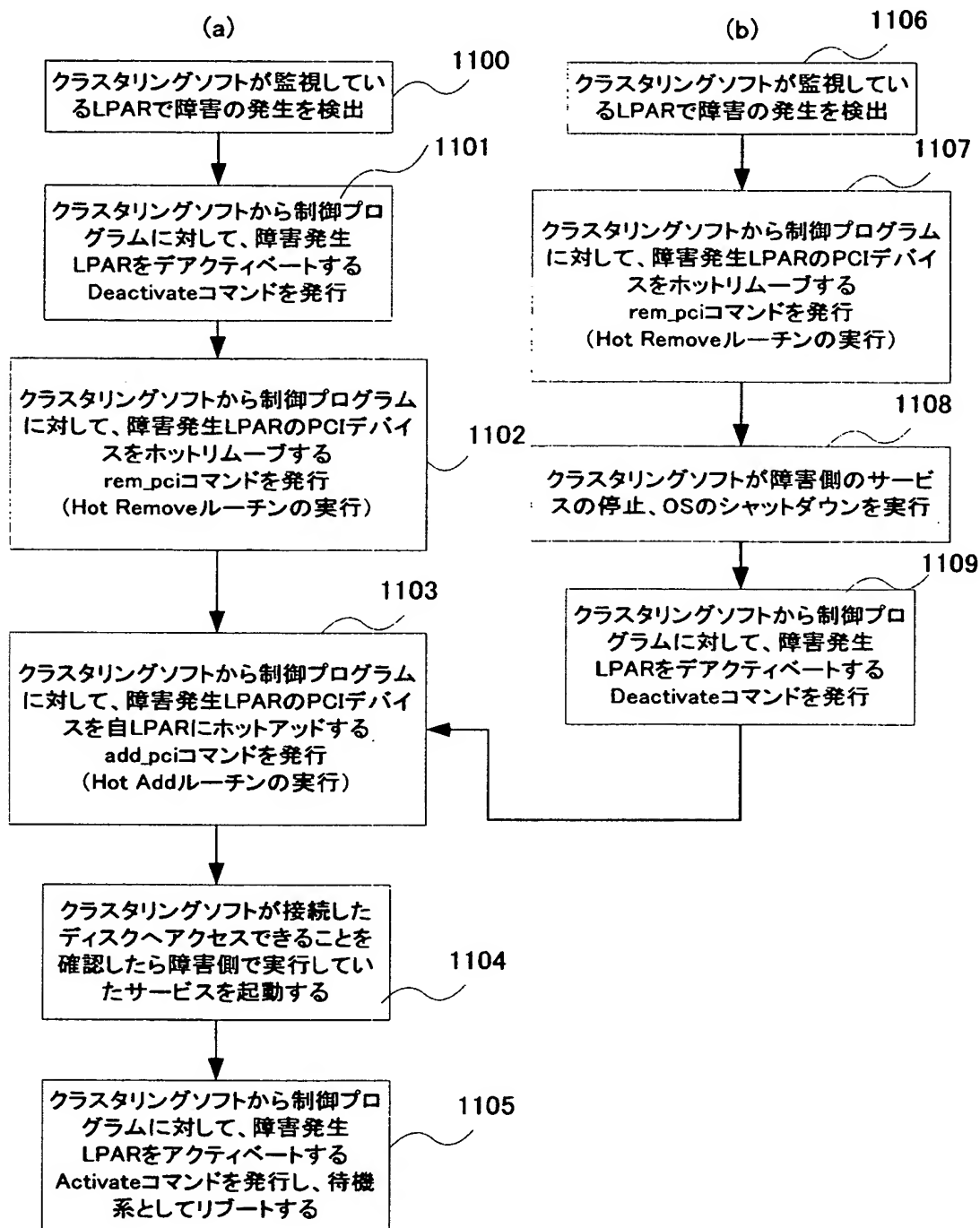
(a)

PCI bus#	BASE	SIZE	OFFSET
0	0	24GB	+0

(b)

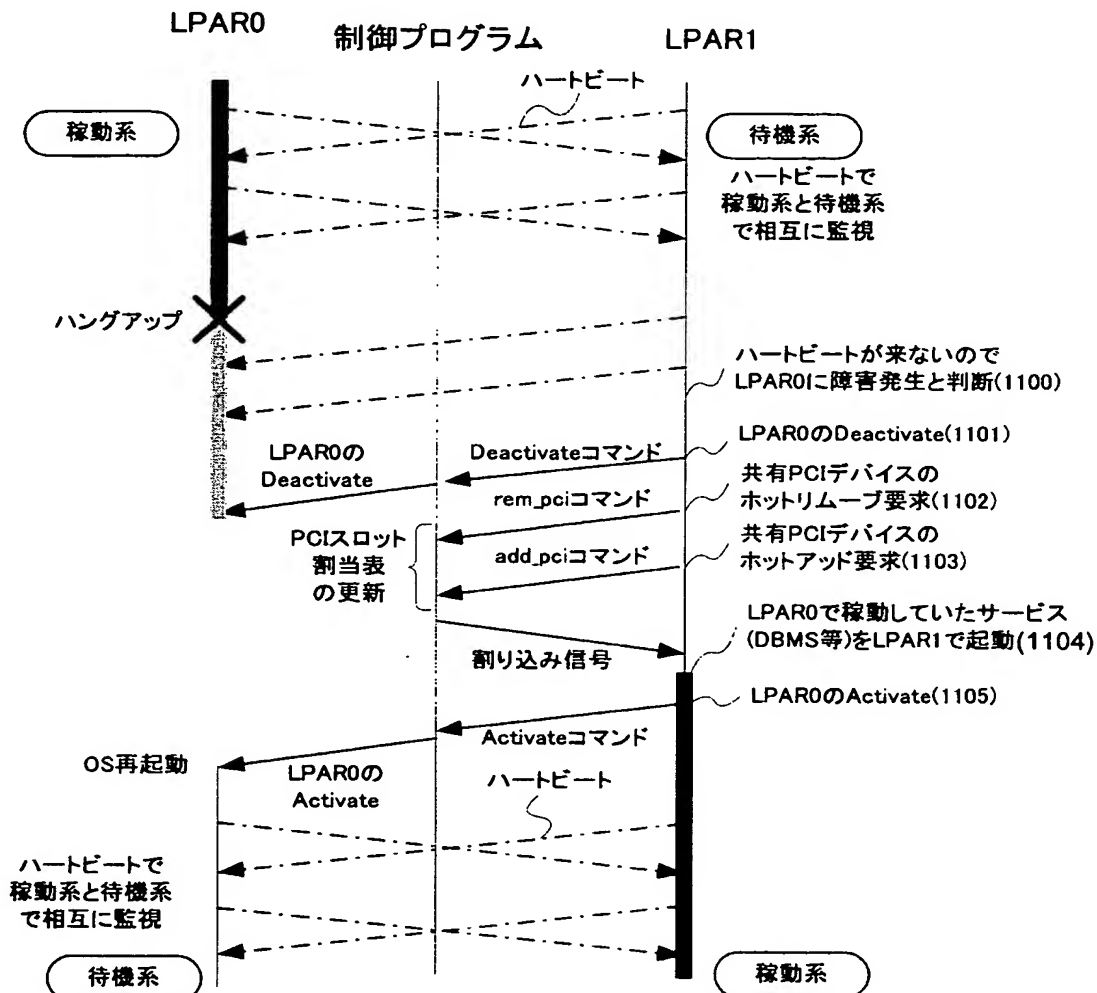
PCI bus#	BASE	SIZE	OFFSET
0	0	32GB	+32GB

【図 9】



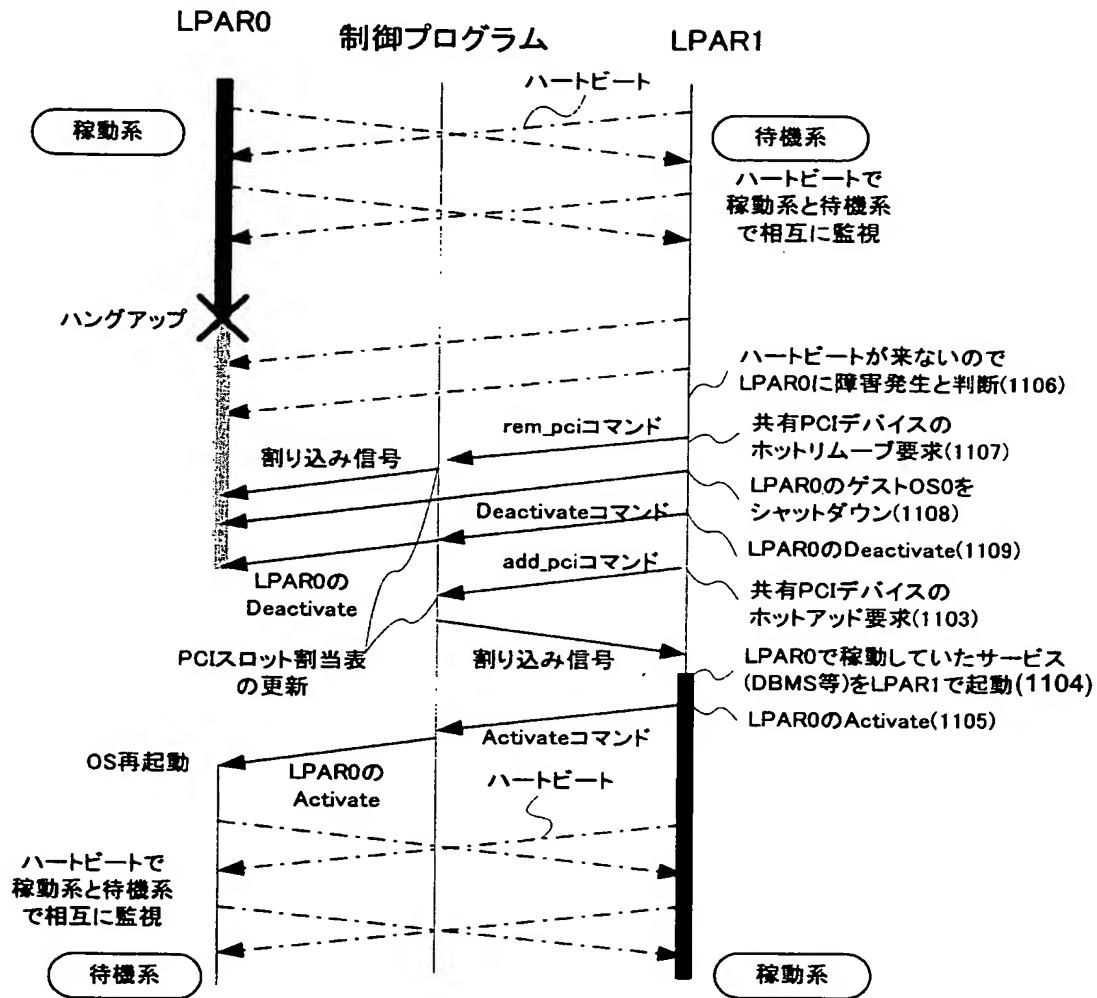
【図10】

(a)

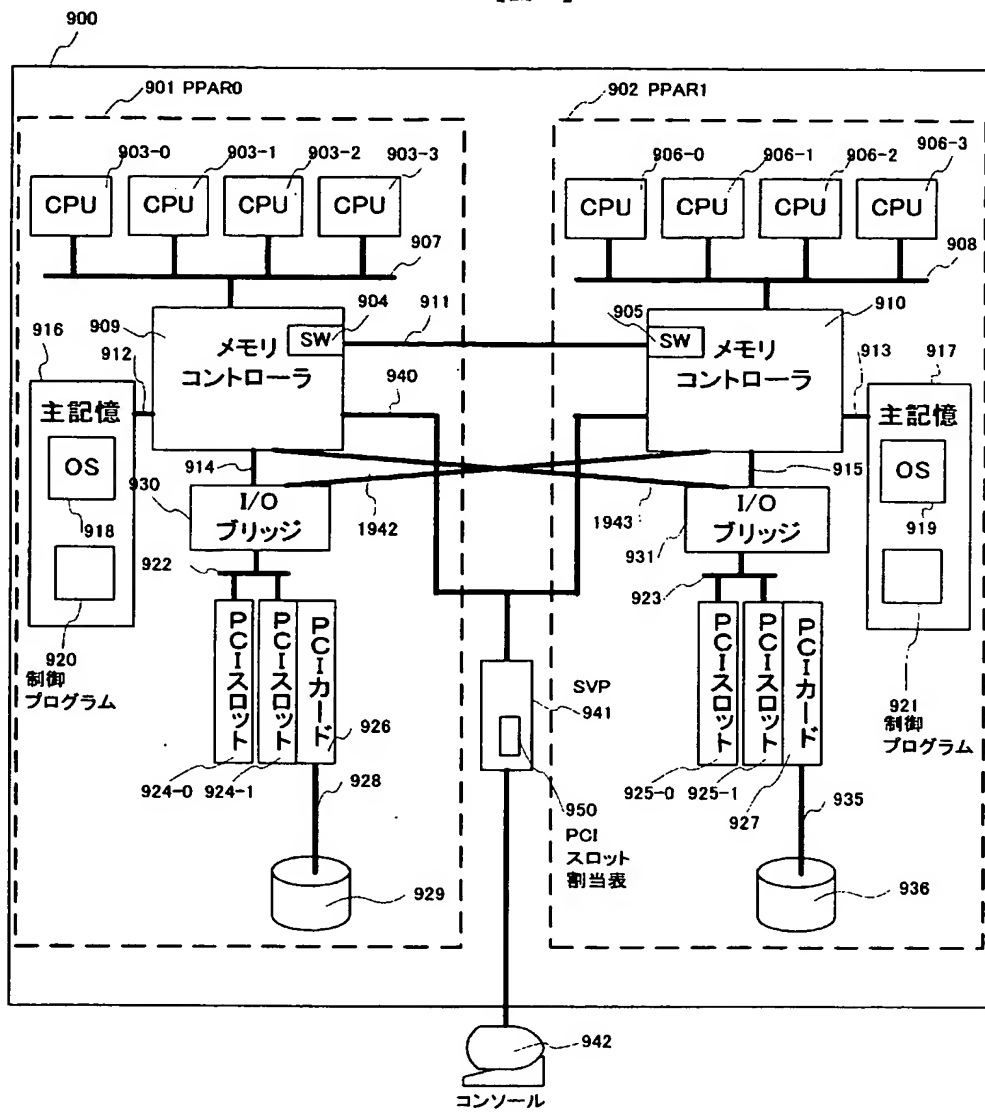


【図 11】

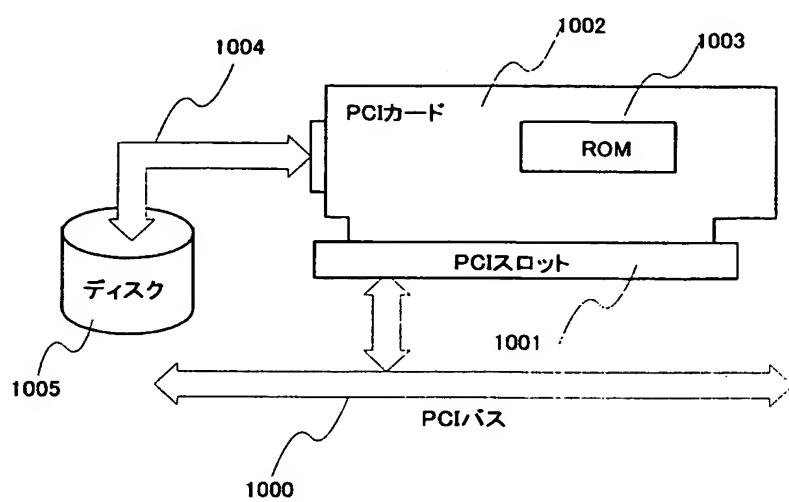
(b)



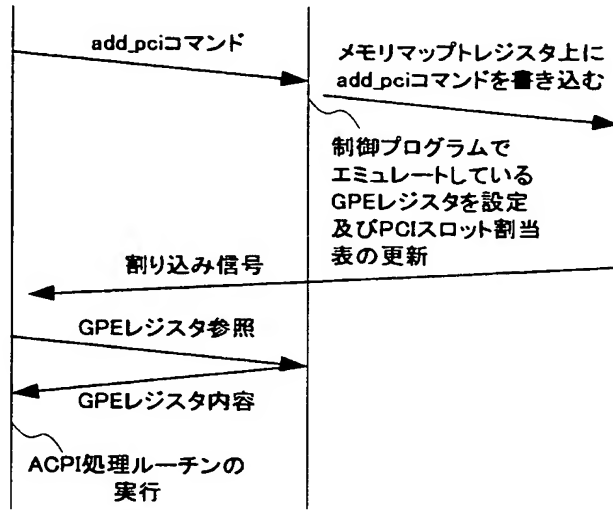
【図 12】



【図 13】



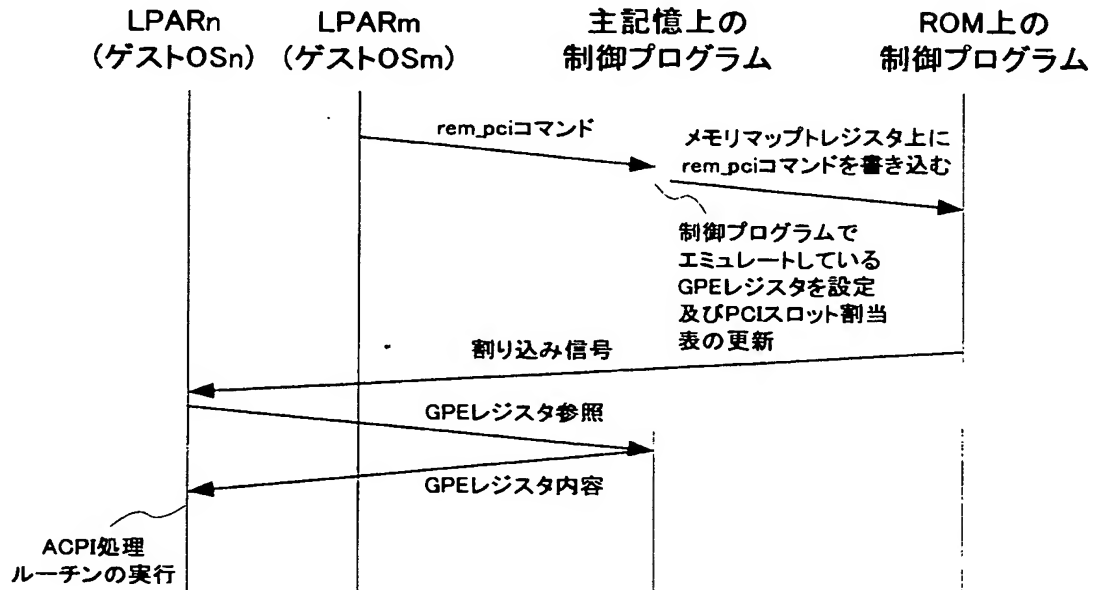
【図 14】

ホット・アッドLPARn
(ゲストOSn)主記憶上の
制御プログラムROM上の
制御プログラム

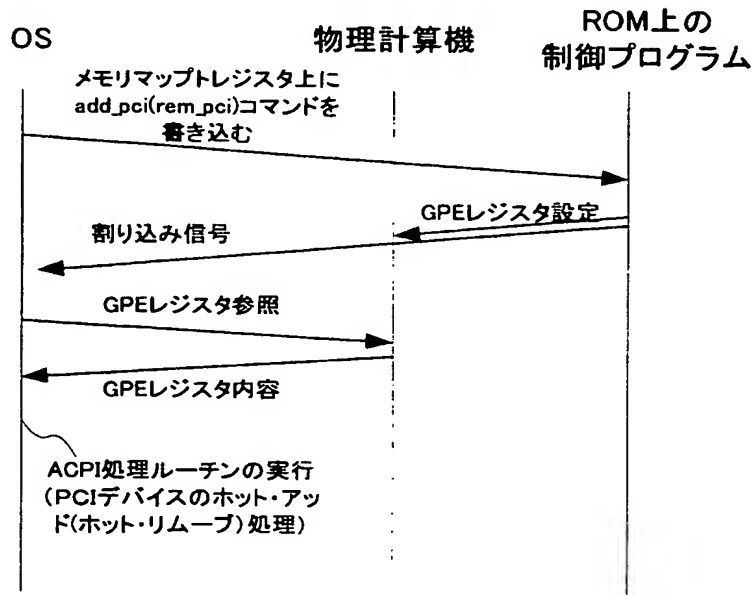
【図 15】

ホット・リムーブ

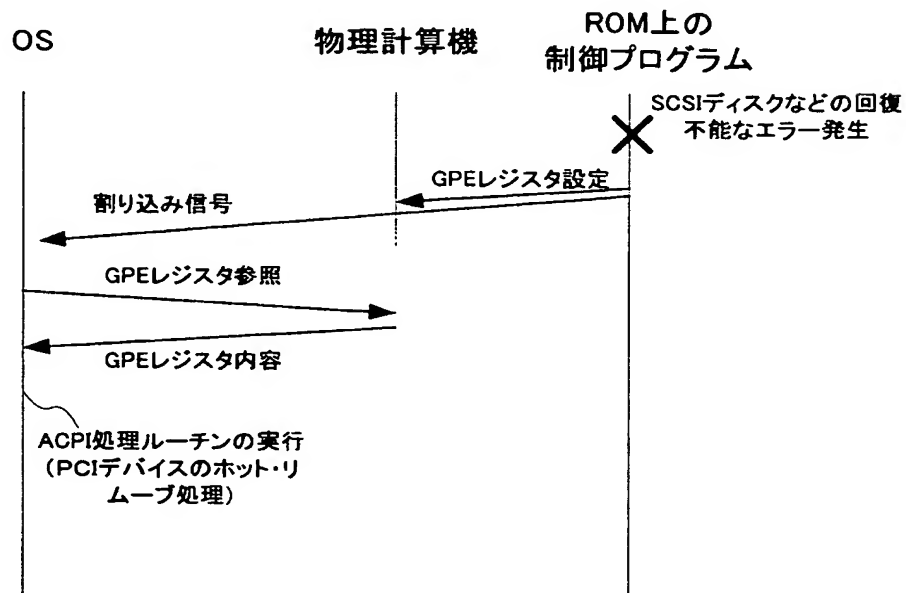
(LPARmからLPARnに割り当てられているPCIデバイスをホット・リムーブする)



【図 16】

ホット・アッド (ホット・リムーブ)

【図 17】

ホット・リムーブ

【書類名】 要約書

【要約】

【課題】 シングルポートのディスクを現用系サーバに接続し、フェールオーバー時には、待機系サーバにこのディスクを接続する。

【解決手段】 クラスタリングソフト 1 0 4 から発行した `add_pci` コマンドにより、制御プログラム 1 0 7 が P C I スロットの割当を変更し、かつ割り込み信号を L A P R 1 (1 0 2) に発行することで、L A P R 1 (1 0 2) のゲスト OS 上で A C P I 処理ルーチンがディスク装置 1 1 2 を含めた P C I カード 1 1 1 をホット・アッドする。

【選択図】 図 2

特願 2 0 0 3 - 0 4 0 2 3 2

出 願 人 履 歴 情 報

識別番号

[0 0 0 0 0 5 1 0 8]

1. 変更年月日

1 9 9 0 年 8 月 3 1 日

[変更理由]

新規登録

住 所

東京都千代田区神田駿河台 4 丁目 6 番地

氏 名

株式会社日立製作所